OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Semester 1 Examinations 2023-2024

Course Instance Code(s)	4BCT1, 4BDS1, 4BMS2, 4BS2, 1MECE1, 1MEME1, SPE, 1EM1, 1AO2, 1OA3, 4BP4, 1MEB1, 1SPS1, 1SPS2
Exam(s)	Fourth BSc in Computer Science & Information Technology Fourth Bachelor of Arts with Data Science Fourth Bachelor of Science (Mathematical Science) Fourth Bachelor of Science (Hons.) ME in Electronic & Computer Engineering ME in Mechanical Engineering Structured PhD
Module Code(s)	CT4101
Module(s)	Machine Learning
Paper No.	1
External Examiner(s)	Dr. Ramona Trestian
Internal Examiner(s)	Professor Michael Madden *Dr. Frank Glavin

Instructions: Answer any 3 questions out of 4 questions.
All questions carry equal marks (25 marks each).
The total (out of 75 marks) will be converted to a percentage after marking.

.

Duration	2 hours
No. of Pages	5
Discipline(s)	School of Computer Science
Course Co-ordinator(s)	Dr. Colm O'Riordan (BCT), Dr. Nick Tosh (BDS), Prof. Dane Flannery (BMS), Dr. Emma Holahan (BS), Prof. Martin Glavin (MECE), Dr. Noel Harrison (MEME)

Requirements:

Release in Exam Venue	Yes [X]	No [ ]
MCQ Answersheet	Yes [ ]	No [X]
Handout	None	
Statistical/ Log Tables	None	
Cambridge Tables	None	
Graph Paper	None	
Log Graph Paper	None	
Other Materials	None	
Graphic material in colour	Yes [ ] No [ ]	

PTO

Question 1 (25 marks)

Part (a)

Below is part of a dataset for a supervised machine learning task.

Anyone for Tennis?					
ID	**Outlook**	**Temp**	**Humidity**	**Windy**	**Play?**
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes

Distinguish between the dependent and independent variables in this dataset.

[1]

What type of supervised learning task will this be used for and why?

[2]

How would this dataset differ if it was being used for the task of clustering?

[1]

Part (b)

In the context of Machine Learning, define what is meant by the terms *Classification*, *Regression,* and *Clustering*. Your definition for each should include a mention of supervised/unsupervised learning, a specific application, and brief description of an algorithm which may be used to learn a suitable model in each case. You should choose a *different* algorithm for each task (to avoid repetition) since there are some algorithms that can cover multiple tasks.

[9]

Part (c)

A chemometrician that you know has sent you the following email message.

"We are looking to identify chlorinated versus non-chlorinated solvents in our lab by collecting the data with a Raman Spectrometer and using some Machine Learning algorithms. I do not have much experience with this but I would like to try using the k-Nearest Neighbours algorithm to build predictive models for two questions: ***whether the spectra are chlorinated or not*** *(we have examples of some of these); and if it is a chlorinated solvent, we wish to* ***detect the percentage of the chlorinated compound that is present*** *(again, we already have some ground truth examples for this that we have collected).*

Can this algorithm be used for both tasks and why?

[2]

Also, I have heard that there are different ways of measuring "nearness", but I don't really understand what this means. Can you please explain what this means, and provide a description of ***three*** *such measures including their equations?*

[6]

PTO

Finally, since we are using multiple Raman Spectrometer machines, the intensity values can often vary when we are collecting data. Is it possible for us to rescale all the feature values with a certain range? If so, how exactly can we do this. Can you outline two different approaches?"

[4]

Prepare a detailed reply answering all the questions above.

Question 2 (25 marks)

Part (a)

Explain what is meant by the term **hyperparameter** in the context of Machine Learning algorithms. Give an example of a hyperparameter, and its possible values, for an algorithm. You should select any algorithm other than k-NN for this part.

[2]

Briefly discuss how a **grid search** could be used to determine suitable hyperparameter values for a machine learning algorithm.

[2]

Part (b)

As part of your job, you have been presented with a new dataset with 500 examples and you have been asked to perform **5 times 5-fold cross-validation with stratification**. Using an illustration, explain this process, including details of how many cases will be in each fold, how the data in each fold is used, and how many classifiers you will build in total. What does it mean to use stratification? How will you estimate the performance of the final classifier?

[6]

Part (c)

Explain how **Receiver Operating Characteristic (ROC)** curves may be used to compare the performance of different classifiers for a binary classification task. Sketch an example of an ROC curve for a binary classification task as part of your answer, clearly labelling the axes and clearly indicating a point that represents ideal performance.

[5]

Part (d)

The table below presents the results of evaluating a classifier on a test set for a binary classification task.

i. Present the results of the evaluation below in a **confusion matrix** [3]
ii. Calculate the **FNR** of the classifier [2]
iii. Calculate the **TNR** of the classifier [2]
iv. Calculate the **precision** of the classifier [2]
v. In terms of model performance, what would a **low value of FNR and FPR** indicate? [1]

PTO

ID	Target	Prediction
1	yes	yes
2	yes	no
3	no	no
4	yes	no
5	no	no
6	no	no
7	no	yes
8	yes	yes
9	yes	yes
10	no	no
11	no	yes
12	no	no
13	yes	no
14	yes	yes
15	no	no
16	yes	yes

[10]

Question 3 (25 marks)

Part (a)

In the context of datasets, describe the terms **bias** and **variance**. Outline the effect of having high bias versus high variance in terms of overfitting/underfitting using some illustrations to support your discussion. Explain how you would identify if your model was currently over- or underfitting.

[7]

Part (b)

Contrast the use of **under-sampling** versus **over-sampling** for dealing with imbalanced datasets.

[4]

What do we mean when we talk about the "*dimensionality*" of a dataset? Describe what is meant by the **curse of dimensionality** and suggest some solutions to address this.

[3]

Part (c)

Explain the difference between **evaluating** a *regression* model versus evaluating a *classification* model. As part of your explanation, you should outline how the following error metrics are calculated: **MSE**; **RMSE**; **MAE** and provide the full names for each.

[4]

PTO

Part (d)
Discuss the key differences between the **k-means** approach versus the **Hierarchical** approach to clustering.

[5]

Part (e)
Illustrate a simple example of how the **elbow method** can be used to help determine k for the k-means algorithm.

[2]

Question 4 (25 marks)

Part (a)
Explain the following concepts and provide the relevant equations for:

- Entropy
- Information gain

[4]

Part (b)

Consider the following set of 15 letters: **computerscience**

i. Calculate the entropy (in bits) of the letters in this set. Hint: you should treat each unique letter as a different class in your calculations. [4]
ii. Calculate the information gain (in bits) if the set of letters is split into two subsets: a subset containing the consonants, and another subset containing the vowels. [6]

For each part, provide complete calculations, along with detailed comments explaining your reasoning for each calculation step.

[10]

Part (c)
Explain how the **McCulloch and Pitts artificial neuron** works. You should include a labelled diagram of the neuron and all equations necessary to calculate the output of the neuron.

[5]

Part (d)
Explain the differences between the following two types of activation functions in the context of neural networks:

i. Logistic
ii. ReLu

[2]

Part (e)
Outline the differences between Binary Encoding and One-hot Encoding by providing a simple example of each. Under what circumstances would these encodings be necessary?

[4]

END